What is claimed is:

1.   An apparatus for recognizing a biological named
entity from biological literature based on united medical
language system (UMLS), comprising:

a resource construction unit for receiving metathesaurus
from the UMLS and constructing a concept name database, a
single name database and a category keyterm database, which
are language resources to be used to recognize a named entity;

a rule collection unit for receiving each concept name
stored in the concept name database, extracting features of
each of the concept names by using data stored in the single
name database and the category keyterm database, and
constructing a rule database by creating a rule used to
recognize the named entity and filtering the rule by using the
extracted features; and

a named entity recognition unit for receiving a
biological literature, extracting nouns and noun phrases that
are candidate named entities, applying the rules stored in the
rule database to the nouns and the noun phrases, and
recognizing the named entities.

2.   The apparatus of claim 1, wherein the resource
construction   unit   extracts   concept   names   from   the
metathesaurus of the UMLS, which is divided according to the
semantic categories, to construct the concept names database,

processes the concept name stored in the concept name database to extract single names and category keyterms, and constructs the single name database and the category keyterm database by using the extracted single names and category keyterms.

3.   The apparatus of claim 1, wherein the rule collection unit extracts the feature of a token constituting each of the concept names stored the concept name database, creates the rules by combining the extracted features, weights the rules, filters the weighted rules with a threshold, and stores the filtered rules in the rule database.

4.   The apparatus of claim 1, wherein the named entity recognition unit extracts the candidate named entities from the literature provided through a literature input unit, extracts the feature of each of the tokens constituting the candidate named entity, creates a rule used to determine the candidate named entity by combining the extracted feature, compares the created rule with the rule stored in the rule database to extract an existing rule suitable for the candidate named entity, applies a weight value of each of the extracted rules and a heuristic used to determine a category of the named entity, determines a final semantic category for the candidate named entity, and recognizing the named entity.

5.   A method for recognizing a biological named entity

from biological literature based on UMLS, the method comprising the steps of:

(a) receiving metathesaurus from the UMLS, extracting concept names, single names and category keyterms, which are language resources to be used to recognize a named entities, and constructing a concept name database, a single name database and a category keyterm database;

(b) extracting features of the concept name by using the language resources stored in each of the databases, constituting a rule for the extracted features, storing the constituted rule in a rule database; and

(c) receiving a literature, extracting features of a candidate named entity, creating a rule used to determine the candidate named entity by combining the extracted features, comparing the created rule with the rules stored in the rule database, and determining a final semantic category by using a result of comparison.

6. The method of claim 5, wherein the step (a) comprises the steps of:

(a-1) mapping information in MRCON table used to describe meaning of each string representing the concept name to information in MRSTY table used to describe a semantic category allocated to each concept name among tables included in the metathesaurus by using a mapping condition, and dividing data stored in the MRCON table according to each

semantic category;

(a-2) extracting values in a string (STR) field of the MRCON table from result of dividing a concept set and storing the extracted values in the concept name database;

(a-3) extracting single names from the concept name database and storing the extracted single names in the single name database; and

(a-4) extracting category keyterms from the concept name database and storing the extracted category keyterm in the category keyterms database.

7. The method of claim 6, wherein in the mapping condition for mapping information in the MRCON table and the MRSTY table, if unique identifier for concept (CUI) of the MRCON table is identical to CUI of the MRSTY table, only data that the value of a language of term (LAT) field is "ENG" among the data in the MRCON table are divided into different sets from one another according to a value corresponding to unique identifier of semantic type (TUI) of the MRSTY table.

8. The method of claim 6, wherein the step (a-4) comprises the steps of:

calculating distribution in the semantic category where each word constituting the named entity appears most frequently by using the concept names stored in the concept name database; and

filtering the words with a threshold.

9.   The method of claim 5, wherein the step (b) comprises the steps of:

(b-1) extracting the features from each of the concept names stored in the concept name database according to a token; and

(b-2) constituting the rule by combining the tokens whose features are extracted, calculating weight value of the constituted rule, filtering the rules with their weight values, and storing the filtered rules in the rule database.

10.   The method of claim 9, wherein in the step (b-1), the feature of the tokens of each of the concept names stored in the concept name database is extracted using the features of the category keyterm, the single name and a capital letter expression, an alphanumeric, a special character, a preposition or conjunction, which are features defined to reflect characteristics of the biological named entity, and a subtype of each of the features.

11.   The method of claim 9, wherein the step (b-2) comprises the steps of:

receiving the result in which the concept name is tokenized and the features are extracted at the step (b-1), and creating the rules as many as the number of combinations

of subtypes according to the subtypes of the features of the token; and

calculating appearance distribution of the rule in each category on all the created rules, filtering the rules with the threshold, and constructing the rule database.

12. The method of claim 5, wherein the step (c) comprises the steps of:

(c-1) extracting nouns and noun phrases, which are candidate named entities, from the inputted literature;

(c-2) extracting features of each token of a candidate named entity;

(c-3) combining the features extracted from each of the tokens of the candidate named entity, and creating the rule used to determine the candidate named entity;

(c-4) comparing the created rule with the rules stored in the rule database; and

(c-5) determining the final semantic category of the candidate named entity.

13. The method of claim 12, wherein in the step (c-4), existing rules suitable to determine the candidate named entity are extracted an existing rule by comparing the rule used to determine the candidate named entity with the rules stored in the rule database in manners of exact match, partial match and nested match.

14. The method of claim 12, wherein in the step (c-5), the final semantic category of the candidate named entity is determined using weight values of existing rules extracted at the step (c-4) and a heuristic used to determine a category of the named entity, and outputted as a result of recognizing the named entity.